ENHANCED LEXICON BASED MODELS FOR EXTRACTING QUESTION-ANSWER PAIRS FROM WEB FORUM

ADEKUNLE ISIAKA OBASA

UNIVERSITI TEKNOLOGI MALAYSIA

ENHANCED LEXICON BASED MODELS FOR EXTRACTING QUESTION-ANSWER PAIRS FROM WEB FORUM

ADEKUNLE ISIAKA OBASA

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MARCH 2016

To my beloved family

ACKNOWLEDGEMENT

First and foremost I will like to thank the almighty Allah Subhanahu wa ta'ala for giving me the strength and wisdom to produce this work. May the peace and blessings of Allah be upon his noble prophet Muhammad (SWA), his family and companions. Next, I wish to thank Kaduna Polytechnic for providing the much needed sponsorship for the programme and Universiti Teknologi Malaysia for giving publication support. I feel highly indebted to my supervisor Professor Dr. Naomie Salim for her guidance, encouragement, support, criticism and very useful suggestions throughout the PhD programme and the thesis preparation.

The support and endurance of my family members deserved being appreciated. I wish to thank my wife and my four children for having to live without me for years. I also wish to thank my aged mother for her incessant prayers. I must not forget to remember all my siblings who always feel concerned about my progress. I will also like to appreciate members of my research group (text mining) for knowledge sharing and collaborations. In addition, members of Soft Computing Research Group (SCRG) especially my lab mates must be appreciated for their useful interaction. At this juncture, I must appreciate the support of Prof. Dr. Siti Mariyam Shamsuddin, the director of UTM Big Data Centre, for her regular big data updates and encouragements. In the same vein, I am grateful to all staff of the Faculty of Computing, UTM, for their kind support throughout my study.

Lastly, I must appreciate Taliat Ola Yusuf, Nicholas Akosu and Adam Isah for taking time out of their tight schedule to generate labels for the three data sets used in the study. Also, I am expressing gratitude to all my friends here in UTM and outside UTM who have in one way or the other assisted, encouraged and motivated me, most especially during the hard times.

ABSTRACT

A Web forum is an online community that brings people in different geographical locations together. Members of the forum exchange ideas and expertise. As a result, a huge amount of contents on different topics are generated on a daily basis. The huge human generated contents of web forum can be mined as question-answer pairs (Q&A). One of the major challenges in mining Q&A from web forum is to establish a good relationship between the question and the candidate answers. This problem is compounded by the noisy nature of web forum's human generated contents. Unfortunately, the existing methods that are used to mine knowledge from web forums ignore the effect of noise on the mining tools, making the lexical contents less effective. This study proposes lexicon based models that can automatically mine question-answer pairs with higher accuracy scores from web forum. The first phase of the research produces question mining model. It was implemented using features generated from unigram, bigram, forum metadata and simple rules. These features were screened using both chi-square and wrapper techniques. Wrapper generated features were used by Multinomial Naïve Bayes to finally build the model. The second phase produced a normalized lexical model for answer mining. It was implemented using 13 lexical features that cut across four quality dimensions. The performance of the features was enhanced by noise normalization, a process that fixed orthographic, phonetic and acronyms noises. The third phase of the research produced a hybridized model of lexical and non-lexical features. The average performances of the question mining model, normalized lexical model and hybridized model for answer mining were 90.3%, 97.5%, and 99.5% respectively on three data sets used. They outperformed all previous works in the domain. The first major contribution of the study is the development of an improved question mining model that is characterized by higher accuracy, better specificity, less complex and ability to generate good accuracy across different forum genres. The second contribution is the development of normalized lexical based model that has capability to establish good relationship between a question and its corresponding answer. The third contribution is the development of a hybridized model that integrates lexical features that guarantee relevance with non-lexical that guarantee quality to mine web forum answers. The fourth contribution is a novel integration of question and answer mining models to automatically generate question-answer pairs from web forum.

ABSTRAK

Forum web adalah komuniti dalam talian yang menyatukan orang ramai yang berada pada kedudukan geografi berbeza. Ahli-ahli forum bertukar idea dan kepakaran. Kesannya, sejumlah besar kandungan mengenai topik yang berbeza dihasilkan setiap hari. Kandungan besar dalam forum web yang didapati daripada manusia boleh dilombong sebagai pasangan soalan dan jawapan (Q&A). Salah satu cabaran utama dalam memperolehi Q&A dari forum web adalah untuk mewujudkan perkaitan yang tepat antara soalan dan calon jawapan. Masalah ini dipersulitkan lagi oleh faktor hingar di dalam kandungan forum web. Malangnya, kaedah sedia ada yang digunakan untuk mengumpul pengetahuan dari forum web mengabaikan kesan hingar pada alat perlombongan dan ini telah menjadikan kandungan leksikal kurang berkesan. Kajian ini mencadangkan model-model yang boleh melombong secara automatik rangkaian soalan dan jawapan dengan ketepatan yang lebih tinggi daripada forum web. Fasa pertama kajian menghasilkan model perlombongan soalan. Ia telah dilaksanakan menggunakan ciri-ciri yang dijana daripada unigram, bigram, metadata forum dan peraturan yang mudah. Ciri-ciri ini telah disaring menggunakan teknik khikuasa dua dan pembalut. Pembalut menjana ciri-ciri yang digunakan oleh teknik Naive Bayes Multinomial untuk membina model. Fasa kedua menghasilkan model leksikal ternormal untuk perlombongan jawapan. Ia telah dilaksanakan dengan menggunakan tiga belas ciri leksikal yang merentasi empat dimensi berkualiti. Prestasi ciri-ciri dipertingkatkan dengan pernormalan hingar, satu proses yang membetulkan hingar ortografik, fonetik dan akronim. Fasa ketiga kajian menghasilkan model ciri hibrid leksikal dan bukan leksikal. Prestasi purata model pengumpulan soalan, model leksikal ternormal dan model hibrid untuk pengumpulan jawapan adalah masing-masing 90.3%, 97.5% dan 99.5% pada tiga set data yang digunakan. Kesemua teknik ini mengatasi teknik sebelumnya yang digunakan dalam domain ini. Sumbangan utama kajian ini adalah pembangunan model pengumpulan soalan yang lebih baik yang mempunyai ciri-ciri ketepatan yang tinggi, lebih spesifik, kurang kompleks dan berkeupayaan untuk menghasilkan ketepatan yang tinggi merentasi genre forum yang berbeza. Sumbangan kedua adalah pembangunan model leksikal ternormal yang berkeupayaan untuk mewujudkan perkaitan yang baik antara soalan dan jawapan. Sumbangan ketiga adalah pembangunan model hibrid yang mengintegrasikan ciri leksikal yang menjamin kerelevanan dengan ciri bukan leksikal yang boleh menjamin kualiti perlombongan jawapan daripada forum web. Sumbangan keempat adalah integrasi baru model pengumpulan soalan dan jawapan untuk secara automatik menjana rangkaian soalan dan jawapan daripada forum web.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

5W1H	-	Who, What, Where, When, Why or How
AC	-	Accuracy
ATC	-	Author is thread creator
BoW	-	Bag-of-Words
CAM	-	Camera
CEN	-	Centroid of a reply to all replies
COR	-	Correlation
COS	-	Post cosine similarity with initial post
CQA	-	Community-based Question Answering
CRF	-	Conditional Random Field
df	-	degree of freedom
Docs	-	Documents
DT	-	Decision Tree
EAT	-	Expected Answer Type
F	-	F-score
FAQ	-	Frequently Asked Question
FN	-	False Negative
FP	-	False Positive
FP	-	False Positive Rate
GAAC	-	Group Average Agglomerative Clustring
HGC	-	Human Generated Contents
IBM	-	International Business Machine
IDF	-	Inverse Document Frequency
IE	-	Information Extraction
INF	-	Information gain
IR	-	Information Retrieval
KBs	-	Knowledge Bases

KL divergence	-	Kullback Leibler divergence
KLD	-	KL divergence of post with initial post
LHS	-	Left Hand Side
LM	-	Language Model
LSP	-	Labelled Sequential Pattern
ML	-	Machine Learning
MNB	-	Multinomial Naïve Bayes
MP	-	Multilayer Perceptron
MT	-	Machine Translation
NB	-	Naïve Bayes
NE	-	Named Entity
NEX	-	No. of exclamation marks
NIST	-	National Institute of Standard and Technology
NL	-	Natural Language
NLP	-	Natural Language Processing
NN	-	Neural Network
NNM	-	No. of noun markers
NNS	-	Noun, plural
NNS	-	No. of non-stop words used
NNW	-	No. of negative words
NP	-	No. of posts in thread
NPC	-	No of posts created by the user
NPV	-	Negative Prediction Value
NPW	-	No. of positive words
NQ	-	Not Question
NQM	-	No. of question marks
NR	-	No. of replies created by the user
NRC	-	No of replies created by the user
NSW	-	No. of stop words
NT	-	No. of threads created by the user
NTC	-	No. of threads created by the user
NTP	-	No. of threads participated in
NW	-	No of words in post

NW+NP+NT+NR	-	Forum Metadata
NWD	-	No. of words
NWH	-	No. of Wh-type words
NWL	-	No. of web links
NYC	-	New York City
ODQA	-	Open Domain Question Answering
OVR	-	Post overlap with initial post
P	-	Precision
PF5	-	Post belongs to the first 5 posts
POS tagger	-	Part of Speech tagger
POS	-	The position of the post in the thread
PPR	-	Probabilistic Phrase Re-ranking
Q	-	Question
Q&A	-	Question and Answer
QA	-	Question Answering
QD	-	Question Detection
QLS	-	Query language similarity with initial post
QM	-	Question Mark
QM+WH	-	Simple Rule
QP	-	Quadratic Programming
R	-	Recall
RB	-	Adverb
RDQA	-	Restricted Domain Question Answering
RIP	-	Ratio of initial post to total posts
RQ	-	Research Question
RRP	-	Ratio of replies to total posts
SMO	-	Sequential Minimal Optimization
SNoW	-	Sparse Network of Winnows
SP	-	Specificity
SPM	-	Sequential Pattern Mining
SPSS	-	Statistical Package for the Social Sciences
SR	-	Simple Rule
SR&FM	-	Simple Rule and Forum Metadata

SR&FM+BoW	-	Integration Approach of Simple Rule, Forum Metadata and Bag of Words
SU	-	Symmetrical Uncertainty
SVM	-	Support Vector Machine
TF	-	Term Frequency
tf-idf	-	term frequency-inverse document frequency
TN	-	True Negative
TP	-	True Positive
TREC	-	Text Retrieval Conference
VBG	-	Verb, gerund or present participle
VBP	-	Verb, non-3rd person singular present
WEKA	-	Waikato Environment for Knowledge Analysis
WH	-	Wh-word type
Wh-queries	-	Who, What, How, Why, etc.
WSD	-	Word Sense Disambiguation
WWW	-	World Wide Web

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The Internet, in particular the World Wide Web has become a huge repository of knowledge. Its paradigm has changed from static to a much more dynamic web. The static web, popularly known as web 1.0 is a web of information sharing and read-only hence limited content. The web nowadays connects people, encourages participation and social interaction. This dynamism gets the web loaded on daily basis with a lot of information. Almost everyone now creates contents on the web. The current search engines are based on keywords and they are capable of returning web links of needed pages. They fail in rendering a user interface that has queries expressed in natural language sentences. It is also not possible for them to return precise answers.

Automated question answering (QA) systems that accept free natural language formulated questions and give in return the exact answer or a small passage that contains the needed information is becoming a necessity (Fan et al., 2008). A type of QA system is frequently asked questions (FAQ). FAQ is a web page (or group of web pages) that provides questions frequently asked by users. These questions are usually about services or different functions of the website. The answers are typically provided with the questions. An FAQ article or list is simply a compilation of Frequently Asked Questions and their answers.

FAQ tags are becoming more popular on websites for many reasons. The main reason is that, they assist in providing support, most commonly, customer support; this in a way reduces repetition of solutions to common problems. Well-designed FAQ pages can assist large organizations in controlling the number of supporting staff that may have to be recruited. FAQ web pages try to assist users in finding answers with less effort. A perfect FAQ page will make people use websites without contacting customer service or help desk. The main constituents of FAQ database are the questions and answers it uses. The focus of this research therefore is to mine automatically question-answer pairs that can be used to populate FAQ database from web forum.

1.2 Problem Background

FAQ's are used to supplement other forms of users' supporting documents such as system's documentation, reference manuals and so on. These other documents are comprehensive and well-structured but they cannot provide independent bits of knowledge that will be needed to solve practical problems. Web forums can give such bits of knowledge but lack coherency, homogeneity and other human content generated problems which make them difficult to explore (Henβ et al., 2012). A high percentage of human generated contents on the web can actually be found in social media websites such as Twitter, web forum, Facebook, and community-based question answering. The consideration of web forum in this research lies in the fact that it has a traditional history. Almost every website has a web forum. The generated content of web forum is much larger than others. There is less segregation in web forum participation compared to other social media. Any member can contribute anything, the restriction is less.

Recently, there appears a new question answering system Community-based Question Answering System (CQA), which is also, referred to as cooperative question answering system, Q&A community, and so on, for example, Baidu zhidao, Yahoo! Answers, Sinaiask, and Stack overflow. In the CQA, users ask question and wait for the answers, and other users who know the answer will give directly his answer to the question. So, in the CQA, people get answers more quickly, effectively and accurately and they do not need to browse lots of web pages before they find the needed answer as it is usually the case in the traditional search engine. And for this reason, CQA develops rapidly, and play important roles in many fields.

Nowadays, CQA is a research focus in the natural language processing (Yan, 2013). The quick response people get from CQA may not at times be fast enough to meet up with spontaneous decision. Hence, there is need for a fast answer retrieval system, a system that will take user's question in its natural form and returns an answer immediately. A Frequently Asked Question (FAQ) system can serve this purpose. In order for FAQ system to meet up with this demand, its question and answer database needs to be populated with natural questions and answers needed by the users of the domain is meant to serve. The current practise mostly is for FAQ manager to populate FAQ database using anticipated questions and answers that users might ask. This practise may make the FAQ database to be populated with Q&A's that are not FAQ. A natural way of doing this is to use the domain's web forum.

Web forums are at times referred to as discussion boards or online forums. They are popular web systems commonly used in different areas such as customer support, community development (Cong et al., 2008), online education and interactive reporting (Hong and Davison, 2009), review comments collection for business intelligence and discovery of expertise networks in online communities (Wang et al., 2009), drug review analysis (Goeuriot et al., 2011) etc. Users of the board exchange ideas while discussing issues and form virtual communities within the discussion board. Multiparty discussions within the forum bring about generation of a large amount of contents on a daily basis on a variety of topics. A web forum contains huge amount of user generated content in form of question-answer pairs (Tawfik Albaham and Salim, 2012). As a result, interest in knowledge discovery and information extraction from web forum has increased in the research community (Hong and Davison, 2009). The large amount of response and the variations of response context lead to the problems of efficient knowledge accumulation and retrieval (Hu et al., 2010). To this end, the myriad of useful information contained in forums can be mined into FAQ.

Several mining techniques have been researched for automatic generation of question-answer pairs from web forum. Mining of web forum questions have been tackled using approaches that ranges from simple rules to complex techniques. The simple rules are the question marks, the 5W1H question words (who, what, where, when, why or how) and modal words (can, will, would etc.). The simple rule approaches are popular methods but have been found to be inadequate for mining web forum questions (Sun et al., 2010; Cong et al., 2008). In a forum, many questions do not end with question mark and most of the statements that contain question words are not questions e.g. "Everybody knows how he behaves". It is also difficult for simple rule to detect imperative questions such as "I wonder if anybody could direct me to her office"

The inadequacies of simple rules called for the implementation of some complex techniques such as sequential pattern mining (SPM) (Cong et al., 2008; Hong and Davison, 2009), n-grams (Sun et al., 2010; Hong and Davison, 2009), and regular expressions (Atkinson et al., 2013). All these approaches could enhance the performance of the simple rules but have their own shortcomings. The SPM requires POS tagger. Accuracy of the method depends on POS tagging. The casual language of forum may affect tagging. Also, the computational effort of the approach may be impractical for large dataset. Higher n-grams are complex to generate and manage. For question mining, differences between question and non-question n-gram must be well established (Sun et al., 2010). (Kwong and Yorke-Smith, 2009) used regular expression to achieve F1 Score of 96% in E-mail domain for interrogative questions. This approach can be built around question words to mine interrogative questions e.g. How [to|do|did|can|could|should|would]. It cannot handle the complex questions that dominate web forum questions.

The problem of question mining from web forum can be attributed to the fact that while it is ease to establish features for questions like question mark, question words etc. we do not have similar features for detecting non-question. It then implies that more effective feature exploration techniques are still required. One of the objectives of this research is to implement bag-of-words and bag-of-bigrams model to generate a lot features for both questions and non-questions. Then, screen these features using dimensionality reduction technique to enhance performance.

The previous work on answer mining from web forum can be classified into supervised, semi-supervised and unsupervised. The supervised methods form the bulk of the methods used so far. In this method, different learning techniques such as Naïve Bayes (Qu and Liu, 2011), Conditional Random Fields (Ding et al., 2008; Kim et al., 2010) and Support Vector Machine (SVM) (Kim et al., 2010; Hong and Davison, 2009; Catherine et al., 2013; Catherine et al., 2012) have been the most popular. The supervised approach has problem of human annotation (labels) of the training and testing samples. The annotations are expensive and time consuming. The semi-supervised (Catherine et al., 2013) and unsupervised (Deepak and Visweswariah, 2014) approaches have just started to come up. These approaches are good for utilizing limited and no answer labels respectively but their accuracy are still too low.

All these approaches require the use of answer detection features. These features can be broadly classified into lexical (textual) and non-lexical (non-textual) features. The features utilize forum post textual content and forum metadata to extract good answers from forums. Textual features are usually used to measure degree of relevance between a document and a query while non-textual features can be employed to estimate the quality of the document (Jeon et al., 2006). Several authors have combined the two to mine good answers from forums as either of the two approaches has been found less effective. A major issue here that is still calling for exploration is the mixture of these textual and non-textual features that will give best performance in terms of precision, recall and F-measures.

In view of the above background, this thesis aimed at developing models using simple but effective approaches to improve the precision, recall, F-score and accuracy of questions and answers detection from web forum by exploiting salient lexicon features.

1.3 Motivation

A preliminary investigation carried out by (Iwai et al., 2010) revealed that about 30 to 40 per cent of inquiries sent to help desk operator can be answered using FAQ. Therefore, more than 30 per cent of inquiries can be solved by providing FAQ pages. FAQ development is an important responsibility of help desk. Companies build up and post FAQ on their web sites to reduce the number of inquiries from users. Users can browse the FAQ pages and clear up their unfamiliar matters before they send emails to help desk. Help desks also analyses inquiry records, constructs question and answer sets, and adds them to their FAQ pages. The task of analysing great deal of inquiries, however, needs a lot of time, hence, the need for automatic FAQ (Iwai et al., 2010). In order to automate FAQ, there has to be a very good source of generating automatically question-answer pairs that will be used to populate the FAQ, hence, the motivation for this research.

Discussion board is mostly being used by people as a problem-solving platform. Web forum members post questions relating to some specific problem, and expect others to provide potential answers. This scenario is depicted in Figure 1.1.

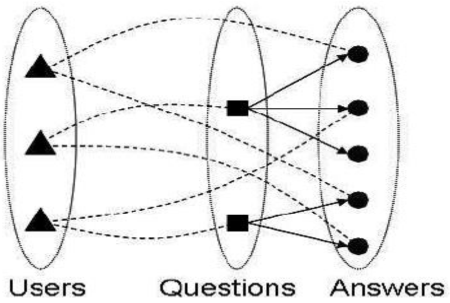

Figure 1.1 Network of Interactions in Forum Connecting Users, Questions and Answers ((Bian et al., 2009)

A number of commercial organizations such as Microsoft, Dell and IBM directly use online forums as problem-solving domain for answering questions and discussing needs raised by customers. (Cong et al., 2008) found that 90% of 40 discussion boards they studied contain question-answering knowledge. By using speech acts investigation on several sampled forums, (Kim et al., 2006a; Kim et al., 2006b) discovered that question answering content is usually the largest type of content on forums. Therefore, mining such content for the benefit of mankind becomes imperative.

The collaborative activities within the forum offer a lot of benefits. In technical forums such as hardware or software forum, a lot of issues such as installing software or hardware, troubleshooting codes, fixing bugs, implementing tools, etc. are being discussed on a daily basis. For non-technical forum like travel, members share their travel experience with others. Good opinions are generated by members for the benefit of other members. It will be highly desirable to mine human knowledge being generated in the forum. Human knowledge mined from forum can be used for numerous applications such as improving question-answer retrieval like frequently asked questions (FAQ); evaluating positive and negative reviews of products' features (Lakkaraju and Ajmera, 2011); or identifying severe technical issues that remain unresolved for a long time (Catherine et al., 2012) etc. These reasons have made mining forum discussions a hot issue in the research community.

Technically, mining question-answer pairs from web forum is an information retrieval (IR) problem. The major issue with this problem is how to establish good relationship between the question posts and reply posts. This appears to be a trivial issue that could be solved using IR techniques. But, it has been found that most of the IR techniques that have proved highly successful prove indifferent in forum posts retrieval. Some good examples are the cosine similarity and KL divergence (Catherine et al., 2012; Raghavan et al., 2010; Wang et al., 2010). In view of this, researchers have started using non-textual features (Catherine et al., 2012) to detect answers in web forum. But, both textual and non-textual features need to be combined to detect answers because an answer should be similar to its corresponding question, and the most intuitive way to achieve this is to establish similarity between them. This scenario calls for an exploration of both the textual and non-textual approaches with a view of striking an effective mix that could guaranty good performance.

In view of all these issues, it is therefore the motive of this research to explore the utilization of more lexical (i.e. textual) and non-lexical (i.e. non-textual) features that could reveal some latent structure to enhance mining of question-answer pairs from web forum.

1.4 Problem Statement

The user of a question answering system is interested in a concise, comprehensive and correct answer, which may refer to a word, sentence, paragraph, image, audio fragment, or an entire document. This research aimed at generating question-answer pairs that could be used to populate FAQ databases from web forum. The research background and motivation have shown some of the contending issues in the research area. The limitations and gaps left by previous research works summarized in Table 1.1 form basis for the key questions to be addressed in this thesis.

Table 1.1: Research Problems Addressed

S/No	Challenges	Sources
1.	Specificity problem	(Sun et al. 2010; Biyani et al. 2015)
2.	Variability of techniques with forum genres	(Hong and Davison 2009; Cong et al. 2008; Liu et al. 2010;)
3.	Complexity of approach	(Cong et al. 2008; Hong and Davison 2009)
4.	Lexical chasm	(Jeon et al. 2006; Catherine et al. 2012; Deepak and Visweswariah, 2015;)
5.	Noise level	(Xue et al., 2011; Wang et al. 2013; HenB et al. 2012)
6.	Relevance problem	(Catherine et al. 2012; Jeon et al. 2006)

⬛ Question mining problems

🟦 Answer mining problems

Hence, the following research questions (RQ) that will be further investigated in this thesis form the premise on which this research stand:

RQ1: How can the lower n-grams be enhanced to increase the accuracy of web forum question post detection?

RQ2: How can the lexical mining tools be enhanced for better performance?

RQ3: What combination of lexical and non-lexical tools will give an optimal performance for answer mining in web forum?

1.5 Research Objectives

The ultimate aim of this study is to develop models using simple but effective approaches to mine question-answer pairs from web forum. Therefore, two hypotheses are set for this study:

a) The use of appropriate classifier with an enhanced "bag of words" using simple rules of question marks, question words, forum metadata and bi-gram can effectively mine questions with good precision, recall and accuracy from web forum.

b) An appropriate combination of both noise-normalized lexical and non-lexical features can effectively mine answers with good precision, recall and accuracy from web forum.

In order to achieve the above stated hypotheses, the following objectives have been set for the study:

i. To investigate the effectiveness of simple rules, forum metadata and lower n-grams for web forum question post detection and propose a model that can improve the accuracy of mining web forum question posts.

ii. To investigate the effect of noise normalization on lexical-based web forum answer mining tools and propose normalized lexical-based model for web forum answer detection.

iii. To propose a hybridized model of normalized lexical and non-lexical features for web forum answer mining task.

1.6 Contribution of the Research

The need to enhance the mining of web forum questions and answers is imperative in view of the knowledge accumulation that is being generated on a daily basis in the forum. The question-answer pairs mined from forum have been found to be useful in so many facets of human lives. The use of web forum question-answer pairs to populate FAQ for service delivery will: provide customer support for common technical problems, minimize the cost of running customer support centres and assist in tackling problem of information overload associated with high volume of Internet transactions. The expected specific contributions of this research are as follows:

i. Selection of important web forum question post classification features can be obtained using chi-square and wrapper techniques.

ii. A novel combination of bag-of-words with simple rules for mining web forum question posts.

iii. Generation of bigram scalable features for mining web forum question posts.

iv. Utilization of some salient features such as centroid of a reply to all replies to significantly influence web forum answer mining performance.

v. Exploration of 13 lexical and 12 non-lexical features for effective mining of answers from web forum.

vi. Evaluation of noise effect on answer mining tools to determine tools that are more susceptible to noise.

vii. Design of a hybrid model of 3 lexical and 2 non-lexical features to effectively mine answers from web forum.

1.7 Research Scope

The following aspects are the scopes of this study:

i. The proposed study will focus on the review of existing literature related to traditional question answering techniques, information retrieval and machine learning. The review will consider both the statistical and computational intelligence approaches that form basis for this research.

ii. The study proposes enhanced BoW model for mining web forum thread using initial post as question thread. An integration of both lexical and non-lexical models is proposed for the mining of a reply post as an answer.

iii. Three publicly available datasets that have been used by some authors that conduct similar research were used to test the proposed method. The datasets are: (a) digital camera dataset, (b) Ubuntu operating system dataset and (c) TripAdvisor dataset. These datasets were carefully chosen to represent less technical, highly technical and non-technical domains (Catherine et al. 2012). This makes it possible to check the performance of the proposed method across different genres.

iv. Mining of web forum question and its corresponding answer is considered as classification problems. The initial post is mined as question post. Clarification questions generated in the reply posts were not considered since they are usually trivial questions. One of the reply posts is mined as the best candidate answer.

v. All the models were implemented using java programming language and weka data mining workbench.

vi. Evaluation of the performance of the models is based on precision, recall, F-score, accuracy and specificity.

vii. Unigram and bigram are considered is this study as lower n-grams.

viii. Lexical mining tools considered in this study are the lexical features and their enhancement is achieved through text normalization.

1.8 Thesis Organization

This thesis consists of seven chapters as follows:

i. Chapter 1: Presents an introduction that describes the concepts of FAQ, and automated question answering system. Thereafter, presents the problem background, motivation, problem statement, objectives, contributions of the study and scopes. The chapter also describes the organization of the thesis.

ii. Chapter 2: Discusses the literature review of the study area. It begins with the fundamental issues in information retrieval and automated question answering system which will be used to realize the workings of the research's methods. Thereafter, a theoretical explanation on the fundamental methods on which the current study is expected to rely is presented. Different techniques of generating question-answer pairs from forum web forum were reviewed. The review focus is primarily on their simplicity and effectiveness to support mining of question-answer pairs that can be used to populate FAQ database.

iii. Chapter 3: Provides the research methodology flow used in this study. It contains the general framework of the research as well as the steps required to carry out the research systematically. The methodological steps discussed comprise of data collection, pre-processing and the processing steps of the research components such as the phases, techniques and tools used.

iv. Chapter 4: Addresses the first objective of the research, which is web forum question post detection using an enhanced bag-of-words approach. Four different approaches were explored in the chapter for the enhancement of bag-of-words, namely, simple rule and forum metadata, bag-of-words, integration of simple rule and forum metadata with bag-of-words and bigram. The chapter also contains experimental results, benchmarking results, and discussions of the approaches.

v. Chapter 5: Covers the second objective of the research. The chapter investigates the effect of noise on answer mining features. In the chapter, forum corpus was normalized against noise. Efficacy of 13 lexical features was explored on normalized and un-normalized forum corpora with a view of finding out how noise will affect the features. Two answer mining models were built for comparisons – normalized lexical model and un-normalized lexical model. The 13-feature normalized model was screened to 5-feature normalized model with a better performance. The chapter covers experimental procedures and relevant discussions of the models. Significant tests were performed to confirm results of the models.

vi. Chapter 6: In this chapter, a hybrid model of lexical and non-lexical was built. This is the third objective of the research. The construction of the model begins with 13 lexical features and 12 non-lexical features. The 25 features were screened using classifier, chi-square and combinatorial analysis to select best mix of 3 lexical and 2 non-lexical for the hybrid. The results of 5-feature hybrid models were evaluated against a main stream approach for web forum answer detection. The chapter covers general overview of the hybrid model as well as experimental procedures, benchmark results and relevant discussions. Significant tests were performed to confirm results of the various models.

vii. Chapter 7: This chapter concludes the thesis by presenting the summaries of all research activities covered in the thesis. The chapter contains brief discussions about the proposed methods, research contributions and feature work.

1.9 Summary

This chapter covers the introductory aspect of the research presented in this thesis. The chapter covers problem background, motivation, problem statement, objectives, expected contribution and the scope of the study. The chapter highlighted the issues in the research area as regards FAQ and mining of web forum question-answer pairs to populate FAQ database.

REFERENCES

Adams, P. H. and Martell, C. H. (2008). Topic detection and extraction in chat. *Proceedings of the 2008 IEEE International Conference on Semantic Computing*,August 4-7, 2008. Santa Clara, CA, USA:IEEE, 581-588.

Agarwal M, Shah R, Mannem P. (2011). Automatic question generation using discourse cues. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. June 24, 2011. Portland.

Aikawa, N., Sakai, T. and Yamana, H. (2011). Community QA Question Classification: Is the Asker Looking for Subjective Answers or Not? *IPSJ Online Transactions.* 4, 160-168. Spriger.

Allam, A. M. N. and Haggag, M. H. (2012). The Question Answering Systems: A Survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS).* 2(3), 211-221.

Arai, K. and Handayani, A. N. (2013). Collaborative Question Answering System Using Domain Knowledge and Answer Quality Predictor. *International Journal of Modern Education and Computer Science (IJMECS).* 5(11), 21-27. MECS.

Atkinson, J., Figueroa, A. and Andrade, C. (2013). Evolutionary optimization for ranking how-to questions based on user-generated contents. *Expert Systems with Applications*. 40(17), 7060-7068. Elsevier.

Bentivogli, L. and Pianta, E. (2000). Looking for lexical gaps. *Proceedings of the ninth EURALEX International Congress, EURALEX 2000.* August 8 - 12, 2000. Stuttgart, Germany, 8-12.

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.* August 15 - 19, 1999. Berkeley, CA, USA, 222-229.

Bernhard, D. and Gurevych, I. (2009). Combining lexical semantic resources with question & answer archives for translation-based answer finding. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.* August 2 - 7 2009. Suntec, Singapore, Volume 2, 728-736.

Bhatia, S., Biyani, P. and Mitra, P. (2014). Summarizing Online Forum Discussions– Can Dialog Acts of Individual Messages Help? *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* October 25-29, 2014. Doha, Qatar, 2127–2131.

Bhatia, S., Biyani, P. and Mitra, P. (2015). Identifying the role of individual user messages in an online discussion and its use in thread retrieval. *Journal of the Association for Information Science and Technology.*
DOI: 10.1002/asi.23373, pg. 1-13. Wiley Online.

Bhatia, S. and Mitra, P. (2010). Adopting inference networks for online thread retrieval. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence.* July 11-15, 2010. Atlanta, Georgia, USA,1300 - 1305.

Bian, J., Liu, Y., Zhou, D., Agichtein, E. and Zha, H. (2009). Learning to recognize reliable users and content in social media with coupled mutual reinforcement. *Proceedings of the 18th international conference on World Wide Web.* April 20 - 24, 2009. Madrid, Spain, 51 - 60.

Biyani, P., Bhatia, S., Caragea, C. and Mitra, P. (2014). Using non-lexical features for identifying factual and opinionative threads in online forums. *Knowledge-Based Systems.* 69 (October 2014), 170-178. Elsevier.

Biyani, P., Caragea, C. and Mitra, P. (2013). Predicting subjectivity orientation of online forum threads. In Gelbukh, A. (Ed.) *Computational Linguistics and Intelligent Text Processing.* (pp. 109-120). Berlin: Springer-Verlag.

Brill, E., Dumais, S. and Banko, M. (2002). An analysis of the AskMSR question-answering system. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.* May 2002. Stroudsburg, PA, USA, 257-264.

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C.-Y., Maiorano, S. and Miller, G. (2010). Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). Available at:

http://www.nlpir.nist.gov/projects/duc/roadmapping.html.

Carpineto, C. and Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR).* 44(1), 1-50. ACM.

Catherine, R., Gangadharaiah, R., Visweswariah, K. and Raghu, D. (2013). Semi-Supervised Answer Extraction from Discussion Forums. *International Joint Conference on Natural Language Processing.* 14-18 October 2013. Nagoya, Japan, 1–9.

Catherine, R., Singh, A., Gangadharaiah, R., Raghu, D. and Visweswariah, K. (2012). Does Similarity Matter? The Case of Answer Extraction from Technical Discussion Forums. *Proceedings of the 24th International Conference on Computational Linguistics. COLING 2012.* December 8 - 15, 2012. Bombay, India, 175-184.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering.* 40(1), 16-28. Elsevier.

Clark, E. and Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. *Procedia-Social and Behavioral Sciences.* 27, 2-11. Elsevier.

Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I. and Sun, Y. (2008). Finding question-answer pairs from online forums. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* July 20 - 24, 2008. Singapore, 467-474.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. New York: John Wiley & Sons.

Deepak, P. and Visweswariah, K. (2014). Unsupervised Solution Post Identification from Discussion Forums. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* June 23-25, 2014. Baltimore, Maryland, USA, 155–164.

Ding, S., Cong, G., Lin, C.-Y. and Zhu, X. (2008). Using conditional random fields to extract contexts and answers of questions from online forums. *Proceedings of Association of Computational Linguistics.* June 2008. Columbus, Ohio, USA, 710 - 718.

Dodiya, T. and Jain, S. (2013). Comparison of Question Answering Systems. In Abraham, A. and Thampi, S. M. (Eds.) *Intelligent Informatics.* (pp. 99-107). Berlin: Springer-Verlag.

Fan, S., Ng, W. W., Wang, X., Zhang, Y. and Wang, X. (2008). Semantic chunk annotation for complex questions using conditional random field. Coling 2008: *Proceedings of the workshop on Knowledge and Reasoning for Answering Questions.* August 2008. Manchester, 1-8.

Flint, L. N. (1917). Newspaper writing in high schools: Containing an outline for the use of teachers. *Publication from the Department of journalism Press in the University of Kansas.*

Gaizauskas, R. J. and Humphreys, K. (2000). A combined IR/NLP approach to question answering against large text collections. *Proceedings of the 6th Content-Based Multimedia Information Access Conference (RIAO-2000).* Paris, France, 1288-1304.

Georgiou, T., Karvounis, M. and Ioannidis, Y. (2010). Extracting Topics of Debate between Users on Web Discussion Boards. *Proceedings of ACM Special Interest Group on Management of Data,(Undergraduate Research Poster Competition).* June 6 - 11, 2010. Indianapolis, Indiana, USA.

Goeuriot, L., Na, J.-C., Kyaing, W. Y. M., Foo, S., Khoo, C., Theng, Y.-L. and Chang, Y.-K. (2011). Textual and informational characteristics of health-related social media content: A study of drug review forums. *Proceedings of Asia-Pacific Conference on Library & Information Education and Practice: Issues, Challenges and Opportunities.* June 22- 24, 2011. Putrajaya, Malaysia, 548 - 557.

Gottipati, S., Lo, D. and Jiang, J. (2011). Finding relevant answers in software forums. *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering.* November 6 - 11, 2011. Oread, Lawrence, Kan, 323-332.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*. 3, 1157-1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter.* 11(1), 10-18.

Hao T, Xie W, Xu F. (2015). A WordNet Expansion-Based Approach for Question Targets Identification and Classification. In Sun, M. et al. (Eds) *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data.* (pp. 333-344). Springer.

Harabagiu, S. M., Moldovan, D. I., Pasca, M., Mihalcea, R., Surdeanu, M., Bunescu, R. C., Girju, R., Rus, V. and Morarescu, P. (2000). FALCON: Boosting Knowledge for Answer Engines. *Text Retrieval Conference (TREC 2000).* November 13 - 16, 2000. Gaithersburg, Maryland, 479-488.

Henβ, S., Monperrus, M. and Mezini, M. (2012). Semi-automatically extracting FAQs to improve accessibility of software development knowledge. *Proceedings of the 34th International Conference on Software Engineering.* June 02 -09, 2012. Zurich, Switzerland, 793-803.

Hermjakob, U. (2001). Parsing and question classification for question answering. *ACL 2001 Proceedings of the workshop on Open-domain question answering-Volume 12.* July 6th. Toulouse, France, 1-6.

Hirschman, L. and Gaizauskas, R. (2002). Natural language question answering: the view from here. *Natural Language Engineering.* 7(04). 275-300. Cambridge.

Hong, L. and Davison, B. D. (2009). A classification-based approach to question answering in discussion boards. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* July 19 - 23, 2009. Boston, Massachusetts, 171-178.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* August 22 - 25, 2004. Seattle, WA, USA, 168-177.

Hu, W.-C., Yu, D.-F. and Jiau, H. C. (2010). A FAQ Finding Process in Open Source Project Forums. *Fifth International Conference on Software Engineering Advances.* August 22 - 27, 2010. Nice, France, 259-264.

Huang, J., Zhou, M. and Yang, D. (2007). Extracting Chatbot Knowledge from Online Discussion Forums. *20th International Joint Conference on Artificial Intelligence (IJCAI-07).* January 6 - 12, 2007. Hyderabad, India, 423-428.

IBM Corp. (2013). *IBM SPSS Statistics for Windows, Version 22.0.* Armonk, NY: IBM Corp.

Iwai, K., Iida, K., Akiyoshi, M. and Komoda, N. (2010). A help desk support system with filtering and reusing e-mails. *2010 8th IEEE International Conference on Industrial Informatics (INDIN).* July 13 - 16, 2010. Osaka, Japan, 321-325.

Jeon, J., Croft, W. B., Lee, J. H. and Park, S. (2006). A framework to predict the quality of answers with non-textual features. *Proceedings of the 29th annual* *international ACM SIGIR conference on Research and development in information retrieval.* August 6 - 10, 2006. Seattle, WA, USA, 228-235.

Kangavari, M., Ghandchi, S. and Golpour, M. (2008). A new model for question answering systems. *World Academy of Science, Engineering and Technology* 42. 536-543.

Khandelwal, S. H. S. (2004). Automatic Topic Extraction and Classification of Usenet Threads. *Natural Language Processing Course Project, Stanford University*, Stanford, California, USA.

Kim, J., Chern, G., Feng, D., Shaw, E. and Hovy, E. (2006a). Mining and assessing discussions on the web through speech act analysis. *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference.* November 5 - 9, 2006. Athens, GA, USA, 1-10.

Kim, J., Shaw, E., Feng, D., Beal, C. and Hovy, E. (2006b). Modeling and assessing student activities in on-line discussions. *Proc. of the AAAI Workshop on Educational Data Mining.* July 16 -17, 2006. Boston, Massachusetts, 1 -8.

Kim, J. W., Candan, K. S. and Dönderler, M. E. (2005). Topic segmentation of message hierarchies for indexing and navigation support. *Proceedings of the 14th international conference on World Wide Web(WWW2005).* May 10 - 14, 2005. Chiba, Japan, 322-331.

Kim, S. N., Wang, L. and Baldwin, T. (2010). Tagging and linking web forum posts. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning.* July 2010. Uppsala, Sweden, 192-202.

Ko, J., Si, L. and Nyberg, E. (2007). A Probabilistic Framework for Answer Selection in Question Answering. *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL).* April 22-27, 2007. Rochester, New York, 524-531.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence.* 97(1), 273-324. Elsevier.

Kolomiyets, O. and Moens, M.-F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences.* 181(24), 5412-5434. Elsevier.

Kumar, N., Srinathan, K. and Varma, V. (2015). Unsupervised Deep Semantic and Logical Analysis for Identification of Solution Posts from Community Answer. *International Journal of Information and Decision Sciences.* 3, 1-26. Report No: IIIT/TR/2015/-1

Kwong, H. and Yorke-Smith, N. (2009). Detection of imperative and declarative question–answer pairs in email conversations. *AI Communications.* 25(4), 271-283.

Labadié, A. and Prince, V. (2008). Intended boundaries detection in topic change tracking for text segmentation. *International Journal of Speech Technology*. 11(3-4), 167-180.

Lakkaraju, H. and Ajmera, J. (2011). Attention prediction on social media brand pages. *Proceedings of the 20th ACM international conference on Information and knowledge management.* October 24 - 28, 2011. Glasgow, United Kingdom, 2157-2160.

Lee, Y., Shamsuddin, S. and Hamed, H. (2008). Bounded PSO vmax function in neural network learning. *Proceedings of the Eighth International Conference on Intelligent Systems Design and Applications, 2008. ISDA'08*. November 26 -28, 2008. Washington, DC, USA: IEEE Computer Society. 474-479.

Li, B., Liu, Y. and Agichtein, E. (2008a). Cocqa: co-training over questions and answers with an application to predicting question subjectivity orientation. *Proceedings of the conference on empirical methods in natural language processing.* October 25 - 27, 2008. Waikiki, Honolulu, Hawaii, 937-946.

Li, F., Zhang, X., Yuan, J. and Zhu, X. (2008b). Classifying what-type questions by head noun tagging. *Proceedings of the 22nd International Conference on Computational Linguistics.* August 2008. Manchester, 481-488.

Li, X. and Roth, D. (2002). Learning question classifiers. *Proceedings of the 19th international conference on Computational linguistics-Volume 1.* August 26 - 30, 2002. Taipei, Taiwan, 1-7.

Lin, C.-J. and Cho, C.-H. (2006). Question pre-processing in a QA system on internet discussion groups. *Proceedings of the Workshop on Task-Focused Summarization and Question Answering.* July 2006. Sydney, 16-23.

Liu, S., Zhong, Y.-X. and Ren, F.-J. (2013). Interactive Question Answering Based on FAQ. In Sun, M. et al. (Eds) *Chinese Computational Linguistics and*

Natural Language Processing Based on Naturally Annotated Big Data. (pp. 73-84). Springer.

Lindberg, D., Popowich, F., Nesbit, J., Winne, P. (2013). Generating Natural Language Questions to Support Learning On-Line. *Proceedings of the 14th European Workshop on Natural Language Generation.* August 8-9, 2013. Sofia, Bulgaria, 105–114,

Lopez, V., Uren, V., Sabou, M. and Motta, E. (2011). Is question answering fit for the semantic web?: a survey. *Semantic Web*. 2(2), 125-155.

Manning, C. D., Raghavan, P. and Schütze, H. (2009). *Introduction to Information Retrieval.* UK: Cambridge University Press, Cambridge.

Mao, J. and Zhu, J. (2012). FAQ Auto Constructing Based on Clustering. *Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering (ICCSEE 2012).* March 23 -25, 2012. Hangzhou, China, 468-472.

Maybury, M. T. (2004). Question Answering: An Introduction. In Maybury, M. T. (Ed.) *New directions in question answering.* (pp. 533-558). Menlo Park: AAAI press.

Mitchell, T. M. (1997). Machine learning. 1997. Burr Ridge, IL: McGraw Hill.

Moldovan, D., Paşca, M., Harabagiu, S. and Surdeanu, M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems (TOIS).* 21(2), 133-154. ACM

Mollá, D. and Vicedo, J. L. (2007). Question answering in restricted domains: An overview. *Computational Linguistics.* 33(1), 41-61. Elsevier.

Moreo, A., Romero, M., Castro, J. L. and Zurita, J. M. (2012). FAQtory: A framework to provide high-quality FAQ retrieval systems. *Expert Systems with Applications.* 39(14), 11525-11534. Elsevier.

Muthmann, K. and Löser, A. (2010). Detecting near-duplicate relations in user generated forum content. In Tang, Y. and Panetto, H. *On the Move to Meaningful Internet Systems: OTM 2010 Workshops.* (pp.698-707). Austria: Springer.

Norvig, P. (2007). How to Write a Spelling Corrector. from <*http://norvig.com/spell-correct.html.> (Accessed 20 November, 2014)*

Pattabiraman, K., Sondhi, P. and Zhai, C. (2013). Exploiting Forum Thread Structures to Improve Thread Clustering. *Proceedings of the ICTIR*

2013International Conference on the Theory of Information Retrieval. September 29 - October 02, 2013. Copenhagen, Denmark, 64 - 71.

Peng, F., Weischedel, R., Licuanan, A. and Xu, J. (2005). Combining deep linguistics analysis and surface pattern learning: A hybrid approach to Chinese definitional question answering. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing.* October 6 - 8, 2005. Vancouver, B. C., Canada, 307-314.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In Scholkopf, B and Burges, C. J. C. (Eds.). .*Advances in kernel methods.* (pp. 185-208). MIT Press.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval.* August 24 - 28, 1998. Melbourne, Australia, 275-281.

Qu, Z. and Liu, Y. (2011). Finding Problem Solving Threads in Online Forum. *Proceedings of the fifth International Joint Conference on Natural Language Processing (IJCNLP 2011).* November 8 - 13, 2011. Chiang Mai, Thailand, 1413-1417.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In Michie, D. (Ed), *Expert systems in the micro electronic age.* (pp. 82 -106). Edinburgh University Press.

Radev, D., Fan, W., Qi, H., Wu, H. and Grewal, A. (2005). Probabilistic question answering on the web. *Journal of the American Society for Information Science and Technology.* 56(6), 571-583.

Raghavan, P., Catherine, R., Ikbal, S., Kambhatla, N. and Majumdar, D. (2010). Extracting problem and resolution information from online discussion forums. *Proceedings of the 16th International Conference on Management of Data (COMAD 2010).* December 8 - 10, 2010. Nagpur, India, 77 - 87.

Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* July 7 -12, 2002. Pennsylvania, USA, 41-47.

Sang, E. T. K., Bouma, G. and De Rijke, M. (2005). Developing offline strategies for answering medical questions. *Proceedings of the AAAI-05 Workshop on*

Question Answering in Restricted Domains. July 9 - 10, 2005. Pittsburgh, PA, USA. 41-45.

Seo, J., Croft, W. B. and Smith, D. A. (2009). Online community search using thread structure. *Proceedings of the 18th ACM conference on Information and knowledge management.* November 2 - 6, 2009. Hong Kong, China, 1907-1910.

Shen, D., Yang, Q., Sun, J.-T. and Chen, Z. (2006). Thread detection in dynamic text message streams. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* August 6 - 10, 2006. Seattle, WA, USA, 35-42.

Shi, L., Sun, B., Kong, L. and Zhang, Y. (2009). Web forum Sentiment analysis based on topics. *Proceedings of the Ninth IEEE International Conference on Computer and Information Technology, (CIT'09).* October 11 - 14, 2009. Xiamen, China, 148-153.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24(4), 35-43.

Smucker, M. D., Allan, J. and Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the sixteenth ACM conference on information and knowledge management.* November 6 - 10, 2007. Lisbon, Portugal, 623-632.

Stoyanchev, S., Song, Y. C. and Lahti, W. (2008). Exact phrases in information retrieval for question answering. Coling 2008: *Proceedings of the 2nd workshop on Information Retrieval for Question Answering.* 24 August, 2008. Manchester, UK, 9-16.

Subramaniam, L. V., Roy, S., Faruquie, T. A. and Negi, S. (2009). A survey of types of text noise and techniques to handle noisy text. *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data.* July 23 - 24, 2009. Barcelona, Spain, 115-122.

Sumit, B., Prakhar, B. and Prasenjit, M. (2012). Classifying User Messages For Managing Web Forum Data. *Proceedings of the Fifteenth International Workshop on the Web and Databases (WebDB 2012),* May 20, 2012. Scottsdale, AZ, USA, 1-6.

Sun, L., Liu, B., Wang, B., Zhang, D. and Wang, X. (2010). A study of features on Primary Question detection in Chinese online forums. *Proceedings of the*

Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD2010). August 10 - 12, 2010. TBD Yantai, China, 2422-2427.

Tang, J., Alelyani, S. and Liu, H. (2014). Feature Selection for Clustering: A Review. In Aggarwal, C. C. and Reddy, C. K. *Data Clustering Algorithms and Applications.* (pp. 30 - 55). US: CRC Press.

Tawfik Albaham, A. and Salim, N. (2013). Online Forum Thread Retrieval using Pseudo Cluster Selection and Voting Techniques. In Chang, R. S. et al. (Eds) *Advances in Intelligent Systems and Applications* (pp. 297 - 306). Springer.

Tomuro, N. (2003). Interrogative reformulation patterns and acquisition of question paraphrases. *Proceedings of the second international workshop on Paraphrasing-Volume 16.* July 11, 2003. Sapporo, Japan, 33-40.

UCLA, (2015). SPSS Annotated Output T-test. UCLA:Institute for digital research and education from
<http://www.ats.ucla.edu/stat/spss/output/Spss_ttest.htm>
(accessed December 26, 2015)

Voorhees, E. M. (2004). Overview of the TREC 2004 robust retrieval track. *Proceedings of Text Retrieval Conference (TREC 2004).* November 19 - 21, 2014. Gaithersburg, Maryland, 1 - 10.

Wang, B.-X., Liu, B.-Q., Sun, C.-J., Wang, X.-L. and Sun, L. (2013). Thread Segmentation Based Answer Detection in Chinese Online Forums. *Acta Automatica Sinica.* 39(1), 11-20. Elsevier.

Wang, B., Liu, B., Sun, C., Wang, X. and Sun, L. (2009). Extracting Chinese question-answer pairs from online forums.. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics (SMC 2009).* October 11 - 14, 2009. San Antonio, Texas, 1159-1164.

Wang, B., Wang, X., Sun, C., Liu, B. and Sun, L. (2010a). Modeling semantic relevance for question-answer pairs in web social communities. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* July 11 - 16, 2010. Uppsala, Sweden, 1230-1238.

Wang, K. and Chua, T.-S. (2010). Exploiting salient patterns for question detection and question retrieval in community-based question answering. *Proceedings of the 23rd International Conference on Computational Linguistics.* August 23 - 27, 2010. Beijing, China, 1155-1163.

Wang, K., Ming, Z.-Y., Hu, X. and Chua, T.-S. (2010b). Segmentation of multi-sentence questions: towards effective question retrieval in cQA services. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.* July 19 - 23, 2010. Geneva, Switzerland, 387-394.

Wang, X. and Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern Recognition.* 36(10), 2429-2439. Elsevier.

Wu, Y., Hori, C., Kashioka, H. and Kawai, H. (2015). Leveraging social Q&A collections for improving complex question answering. *Computer Speech & Language.* 29(1), 1-19. Elsevier.

Xi, W., Lind, J. and Brill, E. (2004). Learning effective ranking functions for newsgroup search. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* July 25 - 29, 2004. Sheffield, UK, 394-401.

Xu, J., Licuanan, A. and Weischedel, R. M. (2003). TREC 2003 QA at BBN: Answering Definitional Questions. *Text Retrieval Conference TREC.* November 18 - 21, 2003. Gaithersburg, Maryland, 98-106.

Xue, Z., Yin, D. and Davison, B. D. (2011). Normalizing microtext. *Proceedings of the AAAI Workshop on Analyzing Microtext.* August 8, 2011. San Francisco, 74-79.

Yan, X. (2013). Question Understanding and Similarity Computation Method Based on Semantic Analysis. *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*. December 8-10, 2012, Beijing, 699-708.

Yang, C.-Y. (2009). A Semantic FAQ System for Online Community Learning. *Journal of Software.* 4(2), 153-158.

Yen, S. J., Wu, Y. C., Yang, J. C., Lee, Y. S., Lee, C. J. and Liu, J. J. (2013). A support vector machine-based context-ranking model for question answering. *Information Sciences*. 224, 77-87. Elsevier

Yu, H. and Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 conference on Empirical methods in natural language processing.* July 11 - 12, 2003. Sapporo, Jappan, 129-136.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*. 22(2), 179-214. ACM

Zhang, D. and Lee, W. S. (2003). Question classification using support vector machines. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* July 28 - August 01, 2003. Toronto, ON, Canada, 26-32.

Zhou, T. C., Lyu, M. R. and King, I. (2012). A classification-based approach to question routing in community question answering. *Proceedings of the 21st international conference companion on World Wide Web.* April 16 - 20, 2012. Lyon, France, 783-790.

Zweigenbaum, P. (2003). Question answering in biomedicine. *Proceedings Workshop on Natural Language Processing for Question Answering, EACL.* April 2003. Budapest, Hungary, 1-4.